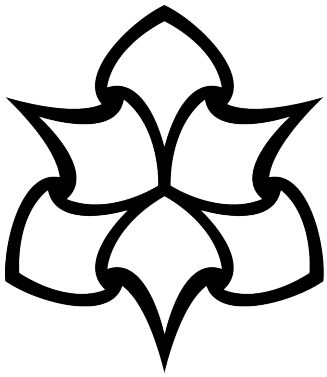

DOI: 10.1002/hrdq.21345

WILEY

LITERATURE REVIEW

Measuring the organizational impact of training: The need for greater methodological rigor

Thomas Garavan[1] | Alma McCarthy[2] | Maura Sheehan[1] | Yanqing Lai[1] | Mark N. K. Saunders[3] | Nicholas Clarke[4] | Ronan Carbery[5] | Valerie Shanahan[6]

[1]Department of International Business, Edinburgh Napier Business School, Edinburgh, Scotland, UK

[2]Discipline of Management, Cairnes School of Business and Economics, National University of Ireland Galway, Galway, Ireland

[3]Department of Management, Birmingham Business School, University of Birmingham

[4]Department of Strategy, Leadership and People, EADA Business School, Barcelona, Spain

[5]Department of Management and Marketing, University College Cork, Cork, Ireland

[6]Global Senior Learning Partner, Squarespace, Dublin, Ireland

Correspondence
Thomas Garavan, Edinburgh Napier Business School, Edinburgh, Scotland, UK.
Email: T.Garavan@napier.ac.uk

All authors contributed equally to this article.

We review the methodological rigor of empirical quantitative studies that have investigated the training and organizational performance relationship. Through a content analysis of 217 studies published in quality journals, we demonstrate significant validity threats (internal, external construct, and statistical conclusion validity) that raise questions about the methodological rigor of the field. Our findings suggest that the time is appropriate for a renewed methodological endeavor to understanding the relationship between training and organizational performance. We make specific recommendations to enhance methodological rigor and generate research findings that will enhance operationalization of theory, help researchers to make inferences about causality, and inform the decision-making of Human Resource Development (HRD) practitioners.

KEYWORDS

methodological rigor, training and organizational performance, validity

1 | INTRODUCTION

In this paper, we review 40 years of quantitative empirical studies that have investigated the training–organizational performance relationship to identify the methodological features of these studies and the extent to which they are subject to validity threats. Training is an important construct in the HRD and learning and development (L&D) disciplines (Bell, Tannenbaum, Ford, Noe, & Kraiger, 2017), and numerous industry-based reports document the considerable investment made by organizations in employee training and development (e.g., Bersin by Deloitte, 2016). In addition, scholars have argued that training enhances organizational performance, including productivity, innovation, customer service quality, and financial performance (Aguinis & Kraiger, 2009; Kim & Ployhart, 2014; Noe, Clarke, & Klein, 2014), yet the evidence base to make these claims is based on a preponderance of cross-sectional research designs that shed

little light on causality. Since 1979, when Miron and McClelland (1979) published the first study on this relationship, the past four decades have witnessed a sustained increase in empirical studies investigating the training–organizational performance relationship, with major growth in published studies since 2010. The extensiveness of past research highlights the importance of training in organizations and the need for researchers to provide practitioners with robust findings on the strength of the relationship, the linking mechanisms, and the boundary conditions explaining the relationship.

While there are many published reviews and syntheses on the topic of training in organizations (e.g., Bell et al., 2017; Noe et al., 2014; Salas, Tannenbaum, Kraiger, & Smith-Jentsch, 2012), these reviews have primarily focused on identifying and reporting key themes and knowledge accumulation on training to date. However, existing reviews seldom engage with the methodological features of studies on the training–organizational performance relationship and the rigor with which research is undertaken. In contrast to prior reviews, our primary aim in this study is to evaluate the methodological characteristics of existing research investigating training and its organizational performance outcomes, specifically to identify the threats to validity that exist in these studies. Given that the training–organizational performance relationship is extensively studied and is central to the arguments that HRD and L&D specialists make to justify investment in training, a major question arises as to the quality of the evidence available on this relationship to date.

Three sets of reasons arise for the need to focus on methodological rigor. First, from the perspective of theory, scholars to date have not always used research designs that reflect the key assumptions of the theories they use to study the relationship. For example, many studies make use of human capital theory (Becker, 1964; Riley, Michael, & Mahoney, 2017) and the resource-based view (Barney, 1991); however, these theories envisage a long-term contribution of investment in human resources to organizational performance. Yet the majority of studies use cross-sectional designs and postpredictive designs (i.e., where respondents provide information on both assessments of current training and their firm's performance at the same time) and therefore do not provide a robust testing of the propositions of the theories used. Wright, Gardner, Moynihan, and Allen (2005) describe these designs as postpredictive because they are actually predicting past performance or performance up to the point of the survey. Similar arguments are made for studies that utilize social exchange theory (Blau, 1964) and behavioral theories (Jackson & Schuler, 1995). Therefore, it is reasonable to conclude that some existing studies do not provide a robust operationalization of the theoretical foundations of these studies.

Second, from an empirical perspective, two important issues arise. First, there is the problem of contextual validity. The majority of studies have been conducted in an Anglo-American context (United States, United Kingdom, Canada, Australia, and New Zealand); therefore, our current understanding of the relationship may not be completely valid given the emergence of Asia-Pacific, Middle Eastern, and African economies. In addition, the majority of studies focus on professional full-time employees, yet the world of employment has changed significantly with the emergence of international workers and the gig economy. This suggests that the context of the training–organizational performance relationship has changed in significant ways, thus suggesting a need to understand the complexities of the relationship. A second empirical reason for analyzing the way in which the training–organizational performance relationship has been investigated concerns the issue of establishing causality. This represents the empirical gold standard of science; however, many existing studies make use of research designs (typically surveys) that do not enable inferences to be made about causality. Wright et al. (2005) highlight that survey designs can never ultimately "prove" cause, and many of what are considered well-designed studies have paid little attention to temporal precedence and/or alternative explanations for the relationship. This issue has also received prominence in the HRD and training literature. For example, both Sitzmann and Weinhardt (2018) and Bainbridge, Sanders, Cogin, and Lin (2017) have drawn attention to the needs for greater methodological rigor in understanding how training and other HRM practices contribute to organizational performance. In the HRD context, Brown and Latham (2018) highlighted the need for both rigor and relevance in HRD research.

Third, from managerial and HRD practice perspectives, it is important to generate valid insights and robust research findings concerning the strength and direction of the relationship between training and organizational performance. Given that the field of HRD focuses on the investigation of L&D processes in workplace settings, it is important that research findings within the field should inform practice in these settings. Thus, an important motivation for this study speaks to recent debates concerning the role of research in generating evidence that is of value in the real world (Brown & Latham, 2018; Gubbins, Harney, van de Werff, & Rousseau, 2018). This discussion suggests that academic

HRD research is moving further away from addressing "real-world" problems that have interest and relevance to practitioners. For research to be relevant to practitioners, it must also be rigorously conducted. Paterson, Harms, and Tuggle (2018) proposed that greater methodological rigor should lead to greater relevance to practitioners. Aguinis et al. (2010) highlighted the concept of customer-centric science and emphasized that careful and rigorous reporting of research results should serve the needs of both academics and practitioners. HRD and L&D scholars are positioned at the theory–practice interface. On the one hand, they generate evidence that can be used by practitioners to make a case for investment in training (Rousseau & Barends, 2011), and on the other hand, they are concerned with the development of a body of knowledge that is robust and answers key theoretical and empirical questions concerning the training–organizational performance relationship (Tharenou, Saks, & Moore, 2007).

Our overarching goal in this paper is therefore to review prior research on the training–organizational performance relationship to illuminate the extent of the validity problem in existing studies and to use the outputs of our analysis to make methodological suggestions to address identified validity threats in future research. In doing so, we seek to enthuse scholars within HRD and L&D to conduct research that achieves the following outcomes. First, scholars should conduct research that provides a strong operationalization of the theoretical perspectives used to formulate hypotheses; second, they should provide a more fine-grained understanding or the training–organizational performance relationship and go further in answering the question of causality; and third, they need to generate findings that will help HRD and L&D practitioners make evidence-based decisions about investment by organizations in training. For the purposes of this paper, validity is defined as the essential trustworthiness of study findings, and scholars have highlighted four categories of validity that are central to methodological rigor (Brutus, Aguinis, & Wassmer, 2013; Casper, Eby, Bordeaux, Lockwood, & Lambert, 2007; Cook & Campbell, 1976). *Internal validity* is concerned with the causality and accuracy of conclusions and is something that plagues much research in the HRD/HRM fields in establishing a relationship between training practices and organizational performance (Bainbridge et al., 2017; Tharenou et al., 2007). *External validity* focuses on the extent to which findings on that relationship are generalizable to different locations, research settings, organizations, employee groups, and across time. *Construct validity* is concerned with the types of measures that are used to operationalize both training and organizational performance, and *statistical conclusion validity* focuses on the extent to which it is possible to make inferences about the training–organizational performance relationship. In quantitative investigations, these dimensions are central to the legitimacy of the field (Bacon, 2016; MacCarthy, Lewis, Voss, & Narasimhan, 2013) of research findings amongst academics and the quality of evidence generated for practitioners (Gelade, 2006).

We make two contributions to the field of HRD and specifically to understanding the training–organizational performance relationship. First, we provide an original overview of existing research on the training–organizational performance relationship in that we discuss key issues related to the validity of the research base. In doing so, we identify methodological issues that have received relatively little attention to date. Second, we advance understanding of the priority validity threats that future researchers should focus on in order to enhance the quality of research findings. For each area of validity, we discuss the research implications of the threats identified and suggest methodological approaches that will decrease or eliminate some of these threats. We structure our paper as follows. We first define the core concepts that underpin the research in this paper. Second, we describe in detail the methodology we used to conduct this study and then present our findings. Finally, we discuss the implications of our findings for methodological rigor and suggest a number of priority recommendations to address causality, contextual validity of studies, the construct validity of the training measure, and greater understanding of linking mechanisms and boundary conditions explaining the relationship.

2 | DEFINING TRAINING AND ORGANIZATIONAL PERFORMANCE

2.1 | Training

Training is defined in different ways in the literature (Bell et al., 2017; Dipboye, 2018), with some definitions emphasizing current knowledge, skill, and ability needs and others focusing on future needs. Training, however, can be

defined as consisting of both "training and development," with the former focused on knowledge, skills, and abilities (KSAs) required for the current job role and the latter focusing on KSAs required for a future role (Garavan, 1995; Kraiger, Passmore, Dos Santos, & Malvezzi, 2014). The future component is conceptualized as development. Training in its narrower sense is sponsored by the organization because it is assumed to have immediate organizational benefits, whereas development may be sponsored by the organization; however, it may also be initiated by employees and without recognition or awareness by the organization. Sitzmann and Weinhardt (2018) argue that the vast majority of training in organizations focuses on what they describe as hard skills or the development of KSAs that are directly applicable to the job. Tharenou et al. (2007), in their meta-analysis of training, focused primarily on these hard skills components and excluded soft skill or development programs. They defined training as "the systematic acquisition and development of the knowledge, skills and attitudes required by employees to adequately perform a job or task and to improve performance" (Tharenou et al., 2007, p. 6). Recent studies of the training–organizational performance relationship have included training focused on enhancing employees' soft skills (Kim & Ployhart, 2014). Therefore, we include in this review studies that reported findings related to training that enhances both current and future KSAs (Berk & Kaše, 2010; Kim & Ployhart, 2014). This definition incorporates training that focuses on the development of generic or soft skills as well as training that takes place in the classroom and on the job (Salas et al., 2012) that is focused on developing hard or skills that are immediately applicable to the job. We selected studies that reported on formal training rather than informal training or training that occurs as part of day-to-day on-the-job experiences, trial and error, and learning by doing (Nelson & Winter, 1982; Nikolova, Van Ruysseveldt, De Witte, & Syroit, 2014). In addition, we only included studies of training conducted in workplace settings.

2.2 | Organizational performance

Organizational performance is conceptualized as a multidimensional construct (Paauwe, 2004) with studies measuring it in different ways. It is the ultimate dependent variable that researchers can use to justify investment in training (Richard, Devinney, Yip, & Johnson, 2009) and includes human resource, operational, and financial performance dimensions (Dyer & Reeves, 1995; Tharenou et al., 2007). However, some studies use the term "organizational effectiveness," which Richard et al. (2009) conceptualize as a broader and more general construct that focuses on internal organizational performance in comparison to external organizational performance measures focused on accounting and financial metrics.

Scholars operationalize organizational performance using objective and subjective measures or a combination of both. The majority of studies utilize subjective measures including, in some cases, a composite index or a single organizational performance item. We define organizational performance to include the three categories proposed by Tharenou et al. (2007): HR-related, operational, and financial. We define human resource outcomes as proximal outcomes such as collective KSAs, motivation, employee turnover, job satisfaction, and organizational commitment (Dyer & Reeves, 1995). We define operational outcomes as distal outcomes comprising labor productivity, innovation, customer service, and customer retention (Jiang, Wang, & Zhao, 2012; Rauch & Hatak, 2016). Finally, we define financial outcomes as comprising three categories: (a) financial performance, (b) product market performance, and (c) shareholder return (Richard et al., 2009). The financial performance category comprises measures of profit, return on assets, and return on investment. Product market performance comprises measures such as sales and market share, and shareholder return includes measures such as total shareholder returns and economic value added. We acknowledge the different approaches taken by scholars concerning this categorization. Rauch and Hatak (2016), for example, did not include HR outcomes as organizational performance outcomes; however, Jiang et al. (2012) in their meta-analysis included HR outcomes in their definition of organizational performance.

3 | METHOD

3.1 | Sample and procedure

We draw on studies published in quality training, HRD, organizational behavior, industrial/organizational psychology, and HRM journals. We examined studies published between 1979 and 2018 to assess the field, and we confined our analysis to articles published in quality journals and specialist journals in the training and HRD fields. We defined a quality journal as those rated 1–4 stars in the Academic Journal Guide, Chartered Association of Business Studies, UK listing (2018). This is an authoritative listing of journal quality. Our starting point for the review was 1979. We utilized this starting point because Tharenou et al. (2007), in the one meta-analysis published to date on the training–organizational performance relationship, identified that year as the starting point for their meta-analysis. We checked to ascertain whether any earlier studies have been published given that the criteria for inclusion in a meta-analysis are more restrictive than those of a methodological review. We searched Business Source Premier, Social Citation Index, and Google Scholar using the following terms: "training and individual outcomes"; "training and organizational outcomes"; or variants of "training and HR outcomes," "training and organizational performance outcomes," "training and organizational effectiveness outcomes," and "training and financial outcomes" to identify relevant articles. We used Google Scholar to search for the most cited articles. We also conducted manual searches of journals that typically publish empirical investigations on the training–organizational performance relationship to ensure that we had captured the relevant articles. Our initial search led to 2,455 articles. To be included in the review, each article was analyzed using three criteria. First, we only included articles that reported empirical findings. We, therefore, excluded papers that were theoretical, conceptual, or literature reviews. This reduced our sample of studies to 1,105 papers. Second, we only included studies conducted in workplace settings, and this further reduced our sample to 756 papers. Third, each study needed to investigate the effects of training on one or more of the three categories of outcome specified by Tharenou et al. (2007)—human resource, organizational, and financial—and to use quantitative methods. This reduced our sample of papers to 217 (the list of papers is presented in Appendix A, Supporting Information). We reviewed the title, abstract, and content of each study against these criteria to determine suitability for inclusion in this review. Our final sample of studies were published in 36 journals, of which the following are examples: *Journal of Organizational Behavior*, *Personnel Psychology*, *The International Journal of Human Resource Management*, *Human Resource Management*, *Human Resource Management Journal*, *Human Resource Development International*, and *Human Resource Development Quarterly*.

3.2 | Coding process

To investigate the four categories of validity, we utilized content analysis (Hoobler & Johnson, 2004; Krippendorff, 2013). Content analysis helps researchers to identify and elaborate on different validity characteristics (Duriau, Rigor, & Pfarrer, 2007). We followed the hierarchically system of codes proposed by Aguinis, Pierce, Bosco, and Muslin (2009) to identify the dimensions to be included in each category of validity.

3.2.1 | Internal validity

We assessed three dimensions of internal validity: (a) the structure of the data (cross-sectional or longitudinal); (b) the research designs used to investigate the training–organizational performance relationship: postpredictive (the measurement of training after the performance period), retrospective (where respondents are asked to recall training practices that existed prior to performance period), contemporaneous (the gathering of concurrent data on training and organizational performance), predictive (the gathering of data on training at one point in time that is related to subsequent organizational performance), or multiple research designs; and (c) the types of relationship investigated (direct, mediated, moderated, moderated mediation).

3.2.2 | External validity

We assessed seven dimensions of external validity: (a) level of analysis of organizational performance (firm, establishment, business unit, multilevel); (b) sample location (North America, Europe, Asia, Africa, Australia/New Zealand, not specified); (c) industry (single industry, multi-industry, not specified); (d) sector (private, public, both, not specified); (e) organization size (specified, not specified); (f) firm/workplace/business unit characteristics (past performance, geographic location, industry or sector, size, age, ownership, competition, number of hierarchical levels, export orientation, diversification, innovation, HR strategy, asset/investment/capital, single or multiple establishment, employee groups, business status, restructuring, level of unionization); and (g) subject-level characteristics (gender, job tenure, education, contract type, working hours, wage levels, age, occupation, race, number of dependents, marital status).

3.2.3 | Construct validity

We assessed the construct validity of both the predictor and dependent variables.

Training

We coded for eight dimensions of the predictor or independent variable: (a) operationalization of the training construct: absolute (the amount of training employees received), proportional (the percentage of workers within an organization trained), content (the type of training provided); emphasis (the perceived importance of the training provided by the organization), effectiveness (the perceived effectiveness of the training provided), or the use of combined measures; (b) training measurement development: existing measure without adaptation, existing measure with adaptation, idiosyncratic (one specifically developed for use in the study), single-item measure, multiple-item measure, binary measure; (c) type of training measure: subjective measures only, objective measures only, subjective and objective measures; (d) number of informants for training measure: single informant, multiple informants, not specified); (e) measurement: reliability evidence for training measure (alpha, interrater, test–retest); (f) measurement: validity evidence of training measure (any content validity evidence, any construct validity evidence, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), discriminant validity, convergent validity); (g) procedural remedies to reduce common method variance (CMV) (where data for both the predictor and dependent variable are obtained from the same person in the same measurement context using the same item context) for the training measure (used, not used); and (h) statistical methods used for CMV for the training measure (used, not used).

Organizational performance

We coded for eight dimensions of the dependent variable: (a) the type of organizational performance measures used (subjective measure only, objective measure only, combined measures); (b) measurement development of the organizational performance variable (existing measures used without adaptation, existing measures used with adaptation, idiosyncratic, single-item measure, multiple-item measure); (c) organizational performance domain measured (human resource, organizational performance financial outcomes, multiple organizational performance outcomes); (d) source of organizational performance measures (same source as training measure, multiple sources, not specified); (e) measurement: reliability evidence of organizational performance measures (alpha, interrater, test–retest); (f) measurement: validity evidence of organizational performance measures (any content validity evidence reported, any construct validity evidence reported, EFA, CFA, discriminant validity, convergent validity); (g) procedural remedies to reduce CMV for the organizational performance variable (used, not used); and (h) statistical methods used for CMV for the organizational performance variable (used, not used).

3.2.4 | Statistical conclusion validity

We coded for nine dimensions of statistical conclusion validity: (a) simple inferential statistics (correlation, *t*-test, chi-square); (b) analysis of statistical relationships (multiple regression, ANOVA and ANCOVA, logistic regression, MANOVA and MANCOVA, canonical correlation, Hierarchical Linear Modeling (HLM), panel analysis, *SEM*, and path

analysis); (c) tests for mediation (Baron and Kenny and alterative models); (d) tests for moderation (MMR); (e) reporting of effect sizes and the magnitude of effect sizes; (f) the reporting of statistical assumption (randomization, independence, measurement level of variable, normality, linearity, and variance); (g) statistical software used to assess relationships (SPSS, Amos, MPlus, LISREL, Stata, not specified); (h) response rate reported (yes, no); and (i) sample size (mean).

3.2.5 | Interrater reliability and validity

Three of the paper's authors were provided with a detailed coding taxonomy developed by the first two authors accompanied by an explanation of each category of validity. Each coder independently coded the data utilizing these coding categories. Our approach is similar to that used by Casper et al. (2007), Hiller, De Church, Murase, and Doly (2011), and Bainbridge et al. (2017). First, the three coders independently coded an initial sample (25) of studies to check for the reliability of coding. Second, we computed the reliability of our coding, made appropriate adjustments, and tightened up the coding taxonomy where necessary. The key challenges we encountered related to the categorization of the training and the organizational performance variables, the categorization of the research design, and the identification of the statically assumptions reported in the paper. Third, following the issuing of new instructions to each coder, we asked a fourth author to code the first set of 25 papers. The first three coders met with the fourth coder to compare coding decisions. We discussed areas of disagreement and explored alternative classification possibilities, and when we reached an agreement, we adjusted the coding taxonomy. The adjustments primarily related to clearly defining the emphasis and effectiveness training variables and broadening our definition of organizational performance to include customer-related outcomes. Where coders had made identical classifications, these consensus codes were recorded in the taxonomy. Each coder then proceeded to code the full set of studies. We calculated agreement between coders for the final coding process using Cohen's kappa level of 0.70 (Brutus et al., 2013). We found the following: Cohen's kappa for each of the four categories in the taxonomy—internal validity (0.90), external validity (0.87), construct validity (0.77), and statistical conclusion validity (0.87).

4 | FINDINGS

4.1 | Internal validity

The key trends that emerge from the analysis on internal validity are summarized as follows.

4.1.1 | Use of cross-sectional designs

Of the studies, 91% used a cross-sectional research design. Cross-sectional designs do contribute to the literature where they are used in the initial phase of investigating novel research questions and potential moderator and mediator hypotheses not previously tested in the literature. They are also useful to help researchers develop new scales and represent a cost-effective way of demonstrating that two or more variables are related to each other. However, cross-sectional designs have limitations in terms of establishing causation, which as we pointed out earlier represents the gold standard in terms of research design. Researchers have expressed concerns about the value of cross-sectional designs to address the fundamental question that underpins organizational investment in training, which is whether training makes a difference to the bottom line. Cross-sectional designs are particularly ineffective when measuring organizational and financial performance outcomes as these types of outcomes require significant time lags to be realized. Only 9% of studies use a longitudinal research design, and they typically measured the training construct at one point in time and used this measure to predict subsequent performance while also controlling for prior or concurrent performance. We encountered significant difficulties in making judgments about the type of research design used in many studies. For example, studies were frequently not precise in describing the timing of training implementation and subsequent measures of performance taken. Studies varied considerably in the time lag

between training and organizational performance. The average time span between the measurement of the training construct and performance was 4.66 years. The longest time was 14 years, and the shortest was 0.5 years. Examples of longitudinal research studies include Kim and Ployhart (2014) and Choi and Yoon (2015). The use of longitudinal designs can help researchers show that changes in training are associated with subsequent changes in organizational performance. This type of design allows a causal type of interpretation to be drawn; however, unless they are an experimental design, the inferences that can be drawn about causality are limited. The limited use of longitudinal designs and the lack of use of experimental designs is a significant limitation of current training–organizational performance research.

4.1.2 | Use of postpredictive designs

The majority (54%) of studies utilize a postpredictive design, which involves the use of organizational performance measurements collected prior to the measurement of the training variable. Wright et al. (2005, p. 412) draw attention to the limitation of postpredictive studies arguing that they "measure HR practices after the performance period, resulting in actually predicting *past* performance." Therefore, while a significant number of studies reported a positive relationship between training and outcomes, it is not possible to make claims about a causal relationship between training and organizational performance due to the overreliance on postpredictive designs. Postpredictive research design involves a single point in time collection of both training and organizational performance data. Researchers typically asked respondents to report current training practices but ask about organizational outcomes up to the point of measurement of the training variable. Examples include Ahmad and Schroeder (2003), Gurbuz and Mert (2011), and Fletcher (2016). A small number of studies use survey methods to gather data on training and archival data to measure outcomes related to past performance (e.g., Beugelsdijkk, 2008; Chen & Huang, 2009). This latter type of study, while interesting, falls into the postpredictive category because the measures of outcomes occurred prior to the measurement of the training variable.

A small number of studies (5%) use "retrospective" designs. These involve asking participants to recall training programs that were in existence prior to the performance period. Examples of studies that use these types of design are Kampkotter and Marggraf (2015) and Zwick (2006). Retrospective research designs are subject to the inaccuracy of recall (Wright et al., 2005) and make it difficult to draw conclusions related to causality. Contemporaneous designs (3%) involve researchers gathering data on training practices and organizational performance data using the same timeframe. Wright et al. (2005) point out that this design is problematic from a causality perspective because the performance data may be gathered both prior to and concurrent with the training practices measure. Predictive designs (13%) investigate whether training implemented at one point in time are related to future organizational performance. Examples of predictive designs include Barrett and O'Connell (2001) and Park and Jacobs (2011). These studies are the most robust in helping researchers draw inferences about causality. Overall, studies demonstrate a positive link between training and organizational performance; however, we can only draw limited conclusions about causality and, for that matter, reverse causality.

4.1.3 | Investigation of direct relationships

The initial stages of the development of a research field typically focus on the measurement of a direct relationship, and as it matures there is a focus on understanding the indirect paths and contingencies that affect the direct relationship. The majority of studies (51%) investigated a direct relationship between training and organizational performance, and researchers continue to investigate a direct relationship; however, the analysis indicates that researchers increasingly investigate linking mechanisms that potentially better explain the link between training and organizational performance and investigated what if- or contingency-type questions. Of the total studies included in our review, 18% reported partially mediated relationships, 14% reported fully mediated relationships, 13% reported moderated relationships, and 4% reported moderated mediation relationships. Therefore, researchers increasingly pay more attention to understanding the processes connecting training to organizational performance and the boundary

conditions that affect the generalizability of direct relationships. The investigation of moderated–mediated relationships is a relatively new statistical method, and we found that a number of recent studies utilized this type of analysis to understand the interaction of linking mechanisms with boundary conditions. However, the use of moderated mediation requires careful operationalization of both the training and organizational performance measures. We found an absence of replication-type studies despite calls for this type of investigation in the HRM, international management, and OB literature (Harzing, 2016).

4.2 | External validity

The following findings emerge on threats to external validity.

4.2.1 | Level of measurement of organizational performance outcomes

We found that the bulk of studies investigated organizational performance at the firm level (74%), with 17% of studies investigating the relationship at the establishment level and 9% at the business unit level. This is an interesting finding because studies that are conducted at the firm level assume that there is little heterogeneity across the firm, whereas studies that utilize a business unit or establishment level of analysis are more likely to capture heterogeneity. This is most likely to be the case in large multinational and multiunit organizations.

4.2.2 | An ethnocentric Anglo-American focus on sample location

We found significant bias in terms of the countries and regions in which data on training and organizational performance is collected. Studies derived samples from five regions, with more than one-quarter from North America and more than one-third from European countries. Of the studies, 27% derived samples from Asia, with the majority of these from China. We found a small number of studies that generated samples from Africa and Australia. There is a significant underrepresentation of samples from Eastern Europe, Latin America, and the Middle East. Therefore, studies have, to date, relied on a small number of countries from which to generate samples, which is a significant threat to external validity and the potential to generalize findings across different countries, cultures, and regions.

4.2.3 | Industry sector and size of firm

The majority of studies report information on industry context. Of studies, 41% were undertaken using single industry samples, and 52% of studies used multiindustry samples. While multiindustry samples help researchers enhance the generalizability of findings, single industry samples help increase measurement precision and allow researchers to capture dimensions of context more effectively. The analysis demonstrates that researchers have not paid attention to the reporting of firm size in empirical investigations. This is not unique to quantitative investigations, with Saunders and Townsend (2016) highlighting that it is also a problem with qualitative studies in general. Of the studies, 40% did not specify the size of the organization when reporting findings or describing the methods used to conduct the study. The lack of attention to the reporting of organization sector and size is particularly problematic, and studies are inconsistent in the way they report organization size: some studies report the mean; others the median; and in other studies, organization size is reported as a log in relation to assets or revenue. These deficiencies in reporting of sector, size, and industry make it difficult for researchers to conduct moderated meta-analysis.

4.2.4 | Organization-, individual-, and subject-level characteristics

Organization- and subject-level characteristics in published studies are not reflective of the diversity of organizations in which training is implemented and the nature of the global workforce in general. There is a major underreporting of both sets of characteristics in existing studies. We found the following trends for the reporting of organization age (20%), ownership (11%), the competitive context (6%), the organization's asset base or level of capital investment (6%), and the level of unionization (12%). There is very poor reporting of individual- or subject-level characteristics. Only 11.5% of studies reported gender, 10% reported job tenure of study respondents, and 9% report education

level. There is a very low level of reporting of employee age (6%), occupation (4%), and race (2%). The majority of studies do not report essential sample characteristics and therefore make it difficult to draw inferences about the generalizability of findings. Even based on the limited reporting of organization- and subject level-characteristics, the samples used in studies do not reflect the diversity of organizations in which training is undertaken and the changing nature of organizations, workforces, and work itself.

4.3 | Construct validity

The following findings emerge on threats to construct validity.

4.3.1 | Operationalization and measurement of the training construct

Clearly defined operationalization of the training construct is a major research design issue. We found four distinct operationalizations of the training construct. Of the studies, 31% operationalize training as a content measure, 7% as an effectiveness measure, 7% as an absolute measure, and 9% as a proportional measure. Of the studies, 20% use a combination of measures. Some of these operationalizations are complex because they involve personal judgments and respondent recall about effectiveness and are therefore potentially subject to random measurement error. Furthermore, measures that focus on effectiveness may be rated more favorably by different categories of study respondents. These errors may lead to the finding of spurious relationships between training and outcomes. Of the studies, 30% utilized idiosyncratic measures exclusively to measure the training construct, 4% used a binary measure, and 13% of studies used a single-item measure. Of the studies, 21% used an existing measure with adaptation, and 13% used an existing measure without adaptation. Overall, many studies create a measure of training that is unique to the study, and the use of single-item measures in controversial and raises important questions about the rigor of measurement of the predictor variable (Fuchs & Diamantopoulos, 2009).

4.3.2 | Use of subjective measures of training and single informants

The use of subjective measures of training and single informants to measure the training construct represents a weakness of published studies. Wall and Wood (2005) highlight the need to secure assessments from two or more persons and the use of the same raters across different organizations. This problem is compounded in multiorganization studies where researchers rely on single informants (e.g., training or HR specialists) who are expected to have knowledge of the training construct. Of the studies, 74% relied on a single informant to provide data on the training construct, and 7% of studies used multiple informants. The majority of studies utilized a subjective measure of training (71%), with 23% of studies utilizing an objective measure such as archival data and 6% using a combination of objective and subjective measures. Researchers criticize studies that rely on single informants due to measurement error issues, low reliability, and statistical inference problems (Sanders & Frenkel, 2011).

4.3.3 | Assessment of reliability, validity, and CMV of training measures

Given the use of both self-reports of training and single-item measurers, there is a low incidence of reporting of reliability. The average α for the training measure was 0.81. A significant number of studies do not pay attention to validity issues. The same issue arises with respect to the reporting of validity evidence due to the use of single-item measures of training. Of the studies, 28% used EFA, 16% used CFA, and 18% report discriminant and 14% convergent validity. Of the studies, 41% did not use procedural remedies to reduce CMV, and 91% of studies did not make use of statistical remedies to address training measure CMV.

4.3.4 | Measurement of organizational performance

Strong research design requires that the measurement of organizational performance variable(s) should be from a different source than that used to measure the training construct. Furthermore, researchers highlight the value of objective measures of organizational performance (Richard et al., 2009). The measurement of organizational performance

is more rigorously measured than the measurement of training. However, research that is more recent highlights the use of subjective measures (Singh, Darwish, & Potočnik, 2016). Of the studies, 58% measured organizational performance using a subjective measure, 32% used an objective measure, and 10% used a combination of subjective and objective measures. The use of objective measures therefore helps ensure that data on organizational performance come from a different source than that of the training measure. Of the studies, 16% used an existing organizational performance measure without adaptation, 36% used a measure of organizational performance with adaptation, 16% used an idiosyncratic measure of organizational performance, 46% of studies used a multiple-item measure of organizational performance, and 26% of studies used a single-item measure of organizational performance. Of the studies, 43% used measures of organizational performance, 24% used measures of financial performance, 23% used measures of human resource outcomes, and 29% of studies use multiple measures of organizational performance.

The collection of data on both training and organizational performance from the same source is problematic (Donaldson & Grant-Vallone, 2002; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). The use of single-source data can have the effect of both inflating and deflating the correlations reported. Of the studies, 61% utilize the same source to measure both the training and organizational performance variables, and in 18% of studies, this dimension was not specified. Therefore, measures of both training and organizational performance are subject to common method bias. These features hamper the extent to which it is possible to infer a relationship between training and outcomes and can result in correlation errors leading to spurious associations.

Given the increased use of multiple items to measure organizational performance, there is a higher incidence of reporting of reliability data (57%). The average α for measures of organizational performance was 0.83. Studies paid less attention to providing evidence of content and construct validity of organizational performance measures. Of the studies, 16% reported evidence of construct validity, and 4% reported evidence of content validity. The reporting of EFA (18%), CFA (12%), discriminant validity (18%), and convergent validity (21%) is low considering researchers make significantly greater use of multiple-item measures of organizational performance. Finally, studies pay little attention to addressing CMV with respect to organizational performance measures. Of the studies, 92% did not report procedural remedies, and 91% of studies do not report statistical remedies to address CMV.

4.4 | Statistical conclusion validity

The following findings emerge on threats to statistical conclusion validity.

4.4.1 | Sample size and response rates

A large sample size helps researchers minimize sampling error. It also affects the extent to which one can generalize. The mean sample size varied depending on the level of analysis of outcomes investigated. The average sample size for firm-level studies is 627; employee's workplace level was 84 employees; and business unit level was 150 employees. Overall, the mean sample size seems appropriate; however, its adequacy depends on how respondents were selected (randomly or convenience), the study purpose, and the data analysis procedures used. In reality, the resources available or the sample size in previous studies frequently determines sample size. However, a variety of data analysis packages, such as MPlus (Muthén & Muthén, 2002), R (Kabacoff, 2017) and Stats (StataCorp, 2013), can be used to determine the appropriateness of the sample size.

The response rate ranged from 22 to 53%, and the average response rate was 43%. We found a lack of clarity and inconsistency in the reporting of response rates. Some studies reported response rates as a percentage of the number sent out, some as a percentage of usable responses, and others as a percentage of those sent out but not deliverable. Studies that use convenience or purposeful samples reported higher response rates than studies using random samples, which reported lower response rates.

4.4.2 | Reporting of effect sizes

We investigated whether studies reported effect sizes, and we analyzed the magnitude of effect sizes found in studies. Both Pek and Flora (2017) and Wilkinson (1999) highlight the importance of the reporting of effect sizes as an important feature of well-conducted research. Overall, we found that many of the earlier studies did not report effect size; however, an analysis of articles from 2010 demonstrates that greater attention is paid to the reporting of effect sizes and the level of significance of effect sizes reported. Effect size was not reported in 48% of studies. In terms of the magnitude of effect sizes reported, we found that the majority of effect sizes reported were small. The distribution of effect sizes using Cohen's (1988) categorization was 42% small (0.20 or more), 33% medium (0.50 or more), and 5% large (0.80 or more). Of the studies, 20% reported an effect size of less than 0.20. Additional analysis of effect sizes indicates that they are significantly lower for the measurement of financial performance compared to operational performance. In addition, they are significantly higher for cross-sectional rather than longitudinal studies and for studies that utilized subjective rather than objective measures of organizational performance.

4.4.3 | Reporting of statistical assumptions

Nimon (2012) highlighted the importance of reporting of statistical assumptions as central to the rigor of quantitative research. We utilized the categorization provided by Nimon (2012) to inform this analysis. Overall, we found very low levels of reporting of statistical significance since 2010. Of the studies, 27% reported on the randomization of the sample data, 14% reported on the independence of data, 26% reported on the measurement level of the training variable, and 33% reported on the measurement level of the organizational performance measure. A slightly larger percentage of studies provided comments or data demonstrating the normality of the data (34%); however, only a small percentage of studies made explicit comments on the linearity of the data (14%) and the issues related to variance (including homogeneity of variance, homogeneity of regression, sphericity, and homoscedasticity) (5%). We did, however, find that these issues were more likely to be reported in studies published in high-ranked journals (4 and 4* journals in the ABS list), and in recent times, the level of reporting of statistical assumptions has improved.

4.4.4 | Use of statistical analysis techniques

The majority of studies reported correlations (78%) followed by *t*-tests (14%) and chi-square tests (1%). To conduct analysis of statistical relationships, studies typically used multiple regression techniques (59%), *SEM* and path analysis (18%), panel analysis (10%), and AVOVA and ANCOVA (10%). In the case of studies that investigated moderation, the majority use MMR, whereas for studies testing mediation, the most common method used was Baron and Kenny's (1986) approach or tests for moderation conducted using *SEM*. In most cases, the software used to conduct analysis is not reported, and the most frequently used packages were SPSS, MPlus, AMOS, and LISREL.

5 | DISCUSSION

This paper set out to investigate the extent of methodological rigor within a homogeneous field of investigation related to the relationship between training and organizational performance. We specifically focused on the extent to which this body of research was subject to internal, external, construct, and statistical conclusion validity threats. Our area of investigation is therefore a very narrow one with distinct boundaries. So what does our review tell us about the state of methodological rigor in training and organizational performance research? Five key trends are apparent: (a) empirical research on the relationship is growing and becoming more international, (b) quantitative methods are the predominant empirical approach, (c) the majority of empirical investigations draws on a very small selection of research methods, and (d) major threats to validity persist within the field. The latter problem is notable in a relatively new field; however, there are also debates concerning more mature fields such as that reviewed here about the lack of precision of measures and methods used in empirical investigations. Rost and Ehrmann (2017), for example, demonstrated that, within the area of management research, there is reporting bias towards win-win results

and Chatterji, Durand, Levine, and Touboul (2016) highlighted significant validity problems with self-report data. Therefore, validity threats are not unique to the training–organizational performance field of investigation.

Overall, the field is characterized by a high degree of methodological conservatism relative to the broader area of management and psychology. Researchers continue to use the same methods that are pervasive within the field despite the significant validity threat problems related to these approaches. In addition, researchers do not often acknowledge these problems and there is a hesitancy to utilize methods that are innovative or more rigorous. These problems highlight a clear need for greater methodological rigor to be a key priority for future research. We suggest that attention to some of the validity threats identified here will help researchers address four core issues: (a) the utilization of methods that will help researchers make inferences about the causal nature of the relationship between training and organizational performance and better operationalization of theories used to generate hypotheses, (b) the generation of samples from unique country and institutional contexts and categories of workers that will help address external or contextual validity issues, (c) greater precision in the measurement of the predictor variable, and (d) the use of more sophisticated research designs to understand boundary conditions and micro-level mechanisms linking training to organizational performance.

5.1 | The pursuit of the gold standard: Demonstrating a causal relationship

To date, researchers have not made sufficient use of research designs that will allow inferences to be drawn about causality. Our analysis highlighted significant threats to internal validity that undermine efforts to achieve this goal. This is, however, a problem that is not unique to training and HRD research, with both Wright et al. (2005) and Bainbridge et al. (2017) highlighting that it is also a problem within strategic human resource management research. However, our analysis highlights that there is a need to utilize research methods that will generate evidence to make a better case for the impact of training on organizational performance. Therefore, there is a case to be made to make greater efforts to utilize longitudinal designs (Ployhart, Weekley, & Ramsey, 2009). They provide an important opportunity but also cause significant challenges for training and organizational performance researchers. The challenge is to collect data on organizational performance sometime after the collection of data on training (Van de Voorde, Paauwe, & Van Veldhoven, 2010) and to collect measures of training and organizational performance at Times 1, 2, and 3. This will allow researchers to make inferences about causality and reverse causality. Training–organizational performance research will be significantly enhanced if researchers track the training investment over time and identify its impacts on organizational performance when training levels are altered or changed. The issue of temporal ordering is central to making inferences about causality; therefore, to do this effectively researchers need to have a minimum of three measurements of both predictor and criterion variables (Ployhart & Vandenberg, 2010). In terms of statistical conclusion validity, this will require the analysis of measurement invariance (Vandenberg, 2002) given that it is difficult to say whether respondents are using the same conceptual frame of reference as they respond to the survey at multiple time periods. It is also important to acknowledge that the use of longitudinal research designs is not without difficulty. Zhu (2012), for example, highlights that longitudinal research designs may suffer from omitted variable bias (Beck, 2011) and endogenous regressors (Hamilton & Nickerson, 2003), and Stritch (2017) highlights the need to investigate variation in data. However, the use of survey methodologies will only go so far in addressing the causality issue.

Experimental designs may be the only effective method in terms of eliminating other alternative explanations for the relationship between training and organizational performance. Studies that have field experiments may be better suited to infer causality. Field experiments are potentially valuable in answering relevant questions about training and outcomes that may be difficult to investigate using other methods (Shadish, Cook, & Campbell, 2002). They can be used to investigate the effects of multiple training conditions. For example, researchers could investigate the performance of high-training versus low-training business units or investigate a strategic training investment choice and its impact on specific outcome metrics. This type of design could help researchers capture the effects of strategic training choices. Field experiments are, however, not the complete answer. They are not particularly useful when researchers wish to understand the mechanisms that explain why training impacted organizational performance. However, they are a significant step in helping researchers explain causality. Field experiments allow researchers to

gather data on outcomes as data that naturally occurs in organizations and allows the independent variable to be manipulated. This situation allows causal inferences to be drawn about the impact of training on organizational performance. Researchers point out that the implementation of field experiments is complex due to the difficulty of finding an equivalent control group. Dehejia and Wahba (2002) proposed propensity score matching (PSM), which helps researchers to match observed characteristics. In the case of training and organizational performance research, the matching can be on issues such as firm size, sector, and industry and technology intensity. Khandker and Samad (2009) proposed the double differences approach, which unlike PSM allows for selection bias regarding unobserved characteristics but assumes that these characteristics do not change over time.

5.2 | Greater attention to external or contextual validity

To date, research on the training–organizational performance relationship is subject to external validity threats or what Ahuja and Novelli (2017) call the problem of contextual validity. This is manifest in a situation where the majority of the research is conducted in Western or developed institutional contexts and is focused on a narrow category of workers. Therefore, much of the research suffers from a generalizability problem. Researchers need to conduct studies in a broader range of countries and generate samples in underrepresented country and institutional contexts, such as the Middle Eastern, Eastern Europe, African, and Latin American countries. We also recommend that researchers generate samples in different industry and sectoral contexts and with firms across micro, SME, and large organizations. For example, there is scope to generate samples in public sector and not-for-profit organizations, and we need more studies within unique industry contexts. There is also a need to study the relationship with different categories of employees. Current research has a strong bias toward investigating white-collar professionals, those who hold full-time jobs, and those who have significant job security working in high-income countries. Bergman and Jean (2016), for example, highlighted the poor representation of low- to medium-skilled employees, temporary workers, and wage earners in industrial–organizational psychology research.

5.3 | Greater precision of measurement of the training variable

Our analysis highlighted significant issues related to construct validity with respect to both the predictor and criterion variables. This problem is demonstrated in the context of training with the overuse of idiosyncratic measures, the use of single-item measures, and the lack of replication of measures in different studies. In the context of training, we found only five studies that used measures of training that were used in two or more previous studies. Researchers have therefore not sufficiently established the construct validity and reliability of published measures across multiple studies. What is also surprising is that well-established measures, such as those found in Fields (2002); the developmental experiences measures developed by Wayne, Shore, and Linden (1997); and components of the learning transfer system (Bates, Holton, & Hatala, 2012), are less frequently used in studies investigating the training–organizational performance relationship. An important challenge in the context of measuring the training construct is the distinction between individual- and organizational-level measures of training. There is a strong tendency towards the use of individual-level perception measures of training related to issues such as effectiveness, importance, and the content of the training, with fewer studies utilizing true organizational-level measures of training, such as the amount of training or the proportion of employees trained.

We recommend the use of archival data to enhance the construct validity of the training measure. Using archival data to measure the training construct may prove valuable because it consists of data gathered in the ordinary course of business without any involvement of a researcher (Spector, Liu, & Sanchez, 2015). Organizations are likely to retain training data for compliance, regulatory, and grant funding purposes. We do, however, acknowledge problems with archival data on training. SMEs and not-for-profit organizations may not gather and maintain accurate, up-to-date training records (Nolan & Garavan, 2016). Furthermore, the training records would not have been created with a particular research question in mind. The lack of match between the data and the question potentially presents internal validity problems. The use of multiple sources for the training construct will provide researchers with better insights into the coverage of the training within an organization.

5.4 | Enhanced understanding of boundary conditions and micro-level mechanisms linking training to organizational performance

An important feature of the growth of a field methodologically is the shift away from the investigation of direct relationships to the investigation of indirect relationships and boundary conditions. We noted that the core mechanisms underlying the training–organizational performance relationship are only beginning to be researched. These linking mechanisms may relate to individual characteristics, leadership, team, organizational and external contextual processes through which training impacts organizational performance. Much of the existing research does not account for the precise mechanisms that link training to organizational performance, and there is a need to jumpstart this line of research by focusing on specific micro linking mechanisms and researching organizational performance outcomes that are proximate to that mechanism and seek out a sample where it may be found. There is a major need to utilize research designs to engage with both contingency and configurational perspectives to investigate the complexities of the training–organizational performance relationship. Scholars in HRM, for example, have highlighted the "black box" problem, and this is equally applicable to the training–organizational performance relationship (Messersmith, Patel, Lepak, & Gould-Williams, 2011). This "black box" is particularly acute in the context of the training–organizational performance relationship where the investigation of boundary conditions is embryonic.

6 | CONCLUSIONS

In this paper, we have conducted a methodological review of the training–organizational performance literature to identify the extent to which it has rigor. We specifically analyzed existing studies to identify threats to internal, external, construct, and statistical conclusion validity. Our analysis of methodological rigor will help researchers make decisions about research designs that more effectively operationalize theories used to investigate the training–organizational performance relationship, utilize methods that enable inferences to be made about causality and reverse causality, and generate a body of research evidence that can be used by practitioners to make decisions about investments in training. We call for renewed vigor and enthusiasm for a significant shift in the way we research that relationship, and argue that old approaches have not served us well in generating evidence that training makes a difference to organizational performance. Rather than simply continue as before, we need to jumpstart the research area by utilizing longitudinal research designs and field experiments, by paying greater attention to the generalizability of research findings by seeking out new contexts in which to conduct research, by paying greater attention to the way we measure training, and finally by researching mediated and moderated relationships. We acknowledge, however, that our review has a number of limitations. First, we focused solely on studies published in the English language and on studies that investigated training as an independent variable. We therefore omitted studies that considered training as a moderator, mediator, or dependent variable. We only included quantitative studies and therefore omitted studies that used qualitative designs. We are, however, confident that enhanced rigor of research on the training–organizational performance relationship will be of benefit to both practitioners and researchers.

ORCID

Thomas Garavan https://orcid.org/0000-0003-2696-7853
Alma McCarthy https://orcid.org/0000-0003-2718-4500
Mark N. K. Saunders https://orcid.org/0000-0001-5176-8317
Ronan Carbery https://orcid.org/0000-0001-8836-2038

REFERENCES

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations and society. *Annual Review of Psychology*, *60*, 451–474.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of organizational research methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, *12*(1), 69–112.

Aguinis, H., Werner, S., Lanza Abbott, J., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*(3), 515–539.

Ahmad, S., & Schroeder, R. G. (2003). The impact of human resource management practices on operational performance: Recognizing country and industry differences. *Journal of Operations Management*, *21*(1), 19–43. https://doi.org/10.1016/S0272-6963(02)00056-6.

Ahuja, G., & Novelli, E. (2017). Redirecting research efforts on the diversification–performance linkage: The search for synergy. *Academy of Management Annals*, *11*(1), 342–390.

Bacon, D. R. (2016). Progress in the legitimacy of business and management education research: Rejoinder to "identifying research topic development in business and management education research using legitimation code theory". *Journal of Management Education*, *40*(6), 700–704.

Bainbridge, H. T. J., Sanders, K., Cogin, J. A., & Lin, C. H. (2017). The pervasiveness and trajectory of methodological choices: A 20-year review of human resource management research. *Human Resource Management*, *56*, 887–913. https://doi.org/10.1002/hrm.21807

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99–120.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Barrett, A., & O'Connell, P. J. (2001). Does training generally work? The returns to in-company training. *Industrial & Labor Relations Review*, *54*(3), 647–662.

Bates, R., Holton, E. F., III, & Hatala, J. P. (2012). A revised learning transfer system inventory: Factorial replication and validation. *Human Resource Development International*, *15*(5), 549–569.

Beck, N. (2011). Of fixed-effects and time-invariant variables. *Political Analysis*, *19*(2), 119–122.

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis*. Chicago, IL: University of Chicago Press.

Bell, B. S., Tannenbaum, S. I., Ford, J. K., Noe, R. A., & Kraiger, K. (2017). 100 years of training and development research: What we know and where we should go. *Journal of Applied Psychology*, *102*(3), 305–323. https://doi.org/10.1037/apl0000142

Bergman, M. E., & Jean, V. A. (2016). Where have all the "workers" gone? A critical analysis of the unrepresentativeness of our samples relative to the labor market in the industrial–organizational psychology literature. *Industrial and Organizational Psychology*, *9*(1), 84–113.

Berk, A., & Kaše, R. (2010). Establishing the value of flexibility created by training: Applying real options methodology to a single HR practice. *Organization Science*, *21*(3), 765–780.

Bersin by Deloitte. (2016). UK *corporate learning factbook 2015: Benchmarks, trends, and analysis of the UK training market*. Retrieved from: www.bersin.com

Beugelsdijkk, S. (2008). Strategic human resource practices and product innovation. *Organization Studies*, *29*(6), 821–847.

Blau, P. M. (1964). *Exchange and power in social life*. New York, NY: John Wiley.

Brown, T. C., & Latham, G. P. (2018). Maintaining relevance and rigor: How we bridge the practitioner–scholar divide within human resource development. *Human Resource Development Quarterly*, *29*(2), 99–105.

Brutus, S., Aguinis, H., & Wassmer, U. (2013). Self-reported limitations and future directions in scholarly reports: Analysis and recommendations. *Journal of Management*, *39*(1), 48–75.

Casper, W. J., Eby, L. T., Bordeaux, C., Lockwood, A., & Lambert, D. (2007). A review of research methods in IO/OB work-family research. *Journal of Applied Psychology*, *92*(1), 28–43.

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal*, *37*(8), 1597–1614.

Chen, C., & Huang, J. (2009). Strategic human resource practices and innovation performance—The mediating role of knowledge management capacity. *Journal of Business Research*, *62*, 104–114.

Choi, M., & Yoon, H. J. (2015). Training investment and organisational outcomes: A moderated mediation model of employee outcomes and strategic orientation of the HR function. *International Journal of Human Resource Management*, *26*(20), 2632–2651.

Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement*, *12*(4), 425–434.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 223–326). Chicago, IL: Rand McNally.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*(1), 151–161.

Dipboye, R. L. (2018). *The Emerald review of industrial and organizational psychology*. London: Emerald.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*(2), 245–260.

Duriau, V. J., Rigor, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies—Research themes, data sources and methodological refinements. *Organizational Research Methods*, *10*(1), 5–34.

Dyer, L., & Reeves, T. (1995). Human resource strategies and firm performance: What do we know and where do we need to go? *International Journal of Human Resource Management*, *6*(3), 656–670.

Fields, D. (2002). *Taking the measure of work: A guide to validated scales for organizational research and diagnosis*. Thousand Oaks, CA: Sage.

Fletcher, L. (2016). Training perceptions, engagement, and performance: Comparing work engagement and personal role engagement. *Human Resource Development International*, *19*(1), 4–26.

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft*, *69*(2), 195.

Garavan, T. N. (1995). HRD stakeholders: Their philosophies, values, expectations and evaluation criteria. *Journal of European Industrial Training*, *19*(10), 17–30.

Gelade, G. A. (2006). But what does it mean in practice? The *Journal of Occupational and Organizational Psychology* from a practitioner perspective. *Journal of Occupational and Organizational Psychology*, *79*(2), 153–160.

Gubbins, C., Harney, B., van de Werff, L., & Rousseau, D. M. (2018). Enhancing the trustworthiness and credibility of human resource development: Evidence-based management to the rescue? *Human Resource Development Quarterly*, *29*, 193–202.

Gurbuz, S., & Mert, I. S. (2011). Impact of the strategic human resource management on organizational performance: Evidence from Turkey. *International Journal of Human Resource Management*, *22*(8), 1803–1822. https://doi.org/10.1080/09585192.2011.565669.

Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization*, *1*(1), 51–78.

Harzing, A. W. (2016). Why replication studies are essential: Learning from failure and success. *Cross Cultural & Strategic Management*, *23*(4), 563–568.

Hiller, N. J., De Church, L. A., Murase, T., & Doly, D. (2011). Searching for outcomes of leadership: A 25-year review. *Journal of Management*, *37*(4), 1137–1177.

Hoobler, J., & Johnson, N. (2004). An analysis of current human resource management publications. *Personnel Review*, *33*, 665–676.

Jackson, S. E., & Schuler, R. S. (1995). Understanding human resource management in the context of organizations and their environments. *Annual Review of Psychology*, *46*(1), 237–264.

Jiang, J., Wang, S., & Zhao, S. (2012). Does HRM facilitate employee creativity and organizational innovation? A study of Chinese firms. *The International Journal of Human Resource Management*, *23*(19), 5024–4047.

Kabacoff, R. (2017). *Bar plots. Quick-R website.* Retrieved from https://www.statmethods.net/graphs/bar.html

Kampkotter, P., & Marggraf, K. (2015). Do employees reciprocate to intra-firm trainings? An analysis of absenteeism and turnover rates. *The International Journal of Human Resource Management*, *26*(22), 2888–2907.

Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2009). *Handbook on impact evaluation: Quantitative methods and practices*. Washington D.C.: The World Bank.

Kim, Y., & Ployhart, R. E. (2014). The effects of staffing and training on firm productivity and profit growth before, during and after the great recession. *Journal of Applied Psychology*, *99*(3), 361–389. https://doi.org/10.1037/a0035408361.

Kraiger, K., Passmore, J., Dos Santos, N. R., & Malvezzi, S. (Eds.). (2014). *The Wiley Blackwell handbook of the psychology of training, development, and performance improvement*. Chichester, UK: John Wiley & Sons.

Krippendorff, K. (2013). *Content analysis: An Introduction to its methodology* (3rd ed.). Chicago, IL: University of Chicago Press.

MacCarthy, B. L., Lewis, M., Voss, C., & Narasimhan, R. (2013). The same old methodologies? Perspectives on OM research in the post-lean age. *International Journal of Operations & Production Management*, *33*(7), 934–956.

Meschi, P. X., & Metais, E. (1998). A socio-economic study of companies through their training policies: New empirical considerations in the French context. *MIR: Management International Review*, *31*(1), 25–48.

Messersmith, J. G., Patel, P. C., Lepak, D. P., & Gould-Williams, J. S. (2011). Unlocking the black box: Exploring the link between high-performance work systems and performance. *Journal of Applied Psychology*, *96*(6), 1105–1118.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620.

Nelson, R. R., & Winter, S. G. (1982). The Schumpeterian tradeoff revisited. *The American Economic Review*, *72*(1), 114–132.

Nikolova, I., Van Ruysseveldt, J., De Witte, H., & Syroit, J. (2014). Work-based learning: Development and validation of a scale measuring the learning potential of the workplace (LPW). *Journal of Vocational Behavior*, *84*(1), 1–10.

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, *3*, 322.

Noe, R. A., Clarke, A. D. M., & Klein, H. J. (2014). Learning in the twenty-first-century workplace. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*, 245–275. https://doi.org/10.1146/annurev-orgpsych-031413-091321

Nolan, C. T., & Garavan, T. N. (2016). Human resource development in SMEs: A systematic review of the literature. *International Journal of Management Reviews*, *18*(1), 85–107.

Paauwe, J. (2004). *HRM and performance: Achieving long-term viability*. Oxford: Oxford University Press.

Park, Y., & Jacobs, R. L. (2011). The influence of investment in workplace learning on learning outcomes and organizational performance. *Human Resource Development Quarterly*, *22*(4), 437–458.

Paterson, T. A., Harms, P. D., & Tuggle, C. S. (2018). Revisiting the rigor–relevance relationship: An institutional logics perspective. *Human Resource Management*, *57*, 1371–1383.

Pek, J., & Flora, D. B. (2017). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, 23(2), 208–225.

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design and analysis of change. *Journal of Management*, *30*(1), 94–120.
Ployhart, R. E., Weekley, J. A., & Ramsey, J. (2009). The consequences of human resource stocks and flows: A longitudinal examination of unit service orientation and unit effectiveness. *Academy of Management Journal*, *52*(5), 996–1015.
Rauch, A., & Hatak, I. (2016). A meta-analysis of different HR-enhancing practices and performance of small and medium sized firms. *Journal of Business Venturing*, *31*(5), 485–504.
Richard, P. J., Devinney, T. M., Yip, G. S., & Johnson, G. (2009). Measuring organizational performance: Towards methodological best practice. *Journal of Management*, *35*(3), 718–804.
Riley, S. M., Michael, S. C., & Mahoney, J. T. (2017). Human capital matters: Market valuation of firm investments in training and the role of complementary assets. *Strategic Management Journal*, *38*(9), 1895–1914.
Rost, K., & Ehrmann, T. (2017). Reporting biases in empirical management research: The example of win-win corporate social responsibility. *Business & Society*, *56*(6), 840–888.
Rousseau, D. M., & Barends, E. G. R. (2011). Becoming an evidence-based manager. *Human Resource Management Journal*, *21*, 221–235.
Salas, E., Tannenbaum, S. I., Kraiger, K., & Smith-Jentsch, K. A. (2012). The science of training and development in organizations: What matters in practice. *Psychological Science in the Public Interest*, *13*(2), 74–101.
Sanders, K., & Frenkel, S. (2011). HR-line management relations: Characteristics and effects. *International Journal of Human Resource Management*, *22*, 1611–1617.
Saunders, M. N. K., & Townsend, K. (2016). Reporting and justifying the number of interview participants in organization and workplace research. *British Journal of Management*, *27*(4), 836–852.
Shadish, W. R., Cook, T., & Campbell, D. T. (2002). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, *60*, 607–629.
Singh, S., Darwish, T. K., & Potočnik, K. (2016). Measuring organizational performance: A case for subjective measures. *British Journal of Management*, *27*(1), 214–224.
Sitzmann, T., & Weinhardt, J. M. (2018). Training engagement theory: A multilevel perspective on the effectiveness of work-related training. *Journal of Management*, *44*(2), 732–756.
Spector, P. E., Liu, C., & Sanchez, J. I. (2015). Methodological and substantive issues in conducting multinational and cross cultural-research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 101–131.
StataCorp, L. P. (2013). *Stata multilevel mixed-effects reference manual*. College Station, TX: StataCorp LP.
Stritch, J. M. (2017). Minding the time: A critical look at longitudinal design and data analysis in quantitative public management research. *Review of Public Personnel Administration*, *37*(2), 219–244.
Tharenou, P., Saks, A. M., & Moore, C. (2007). A review and critique of research on training and organizational-level outcomes. *Human Resource Management Review*, *17*(3), 251–273. https://doi.org/10.1016/j.hrmr.2007.07.004
Van De Voorde, K., Paauwe, J., & Van Veldhoven, M. (2010). Predicting business unit performance using employee surveys: Monitoring HRM-related changes. *Human Resource Management Journal*, *20*(1), 44–63.
Vandenberg, R. J. (2002). Towards a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158.
Wall, T. D., & Wood, S. J. (2005). The romance of human resource management and business performance, and the case for big science. *Human Relations*, *58*(4), 429–462.
Wayne, W. J., Shore, L. M., & Linden, R. C. (1997). Perceived organizational behaviors and their effects on organizational effectiveness in limited-menu restaurants. *Academy of Management Journal*, *40*(1), 82–111.
Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594, 604.
Wright, P. M., Gardner, T. M., Moynihan, L. M., & Allen, M. R. (2005). The relationship between HR practices and firm performance: Examining causal order. *Personnel Psychology*, *58*, 409–446.
Zhu, L. (2012). Panel data analysis in public administration: Substantive and statistical considerations. *Journal of Public Administration Research and Theory*, *23*(2), 395–428.
Zwick, T. (2006). The impact of training intensity on establishment productivity. *Industrial Relations: A Journal of Economy and Society*, *45*(1), 26–46.

AUTHORS' BIOGRAPHIES

Thomas Garavan is Research Professor of Leadership, specializing in leadership development, HRD and leadership, CSR and leadership, and cross-cultural leadership in Edinburgh Napier University Business School. He is Editor of the *European Journal of Training and Development* and Associate Editor of *Personnel Review*.

Alma McCarthy is Professor of Public Sector Management at the National University of Ireland, Galway. Her research interests include public sector leadership and human resource development, training, work-life balance, and 360° feedback. She is on the Editorial Board of *Human Resource Management Journal*, *Human Resource Development Quarterly*, *European Journal of Training and Development*, *the Journal of Managerial Psychology*.

Maura Sheehan is Professor of International Management and Director of the Centre for International Management and Governance Research (ICMGR) at Edinburgh Napier University. She specializes in the relationship between HRM, HRD and organizational performance. She was an immediate past Associate Editor of *Human Resource Development Quarterly (HRDQ)* and *Journal of Organisational Effectiveness: People and Performance*.

Yanqing Lai completed her PhD in Kingston University London, and is currently a research assistant in HRM/Leadership subject group in Edinburgh Napier University. Her research interests mainly lie within strategic human resource management, and strategy and performance in SMEs. She has been published in top-tier journals including *Journal of Business Venturing* and *Human Resource Management Review*.

Mark N. K. Saunders is Professor of Business Research Methods at the University of Birmingham. His research interests include research methods, in particular participant selection and methods for understanding organizational relationships, human resource aspects of the management of change, in particular trust within and between organizations and organizational learning, and small and medium sized enterprise (SME) success.

Nicholas Clarke is Professor of Organisational Behaviour & HRM at EADA Business School Spain. His research focuses on how the quality of work relationships (manager-subordinate relations, team relations, social relations) influence the effectiveness and outcomes of human resource development in organizations including leadership development. He is on the editorial board of *Human Resource Development Quarterly*, *European Journal of Training & Development*, and *Team Performance Management*.

Ronan Carbery is a Senior Lecturer in HRM/HRD at University College Cork. His research interests include career development, talent management, and participation in HRD activities. He is co-editor of the European Journal of Training and Development and serves as an Editorial Advisory Board member on *Human Resource Development Quarterly* and *Human Resource Development International*.

Valerie Shanahan is Global Senior Learning Partner with Squarespace. Previously, she was a Lecturer in Human Resource Management at Dublin City University including lecturing in their international campus in Riyadh, Saudi Arabia. Her research interests include strategic HRD and how learning and development investment impacts organizational performance.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Garavan T, McCarthy A, Sheehan M, et al. Measuring the organizational impact of training: The need for greater methodological rigor. *Human Resource Development Quarterly*. 2019;30: 291–309. https://doi.org/10.1002/hrdq.21345